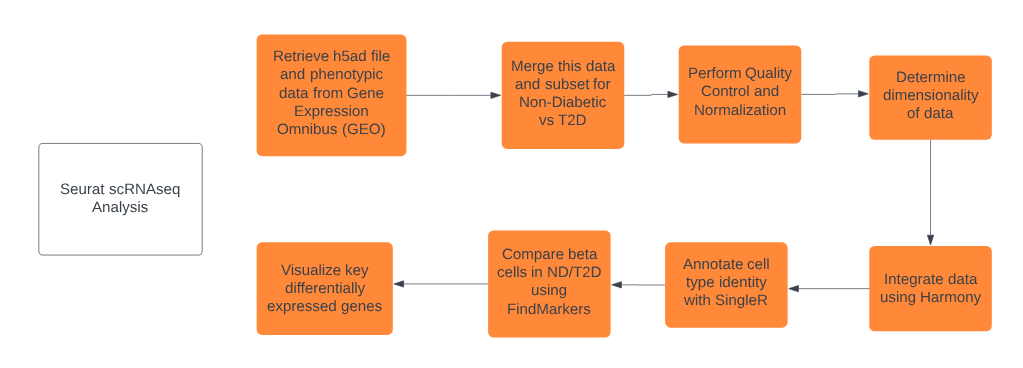
Performing a Harmony integrated analysis of pancreatic islet data

Brandon Thong - bt1194

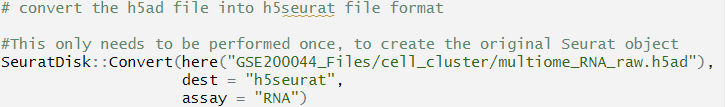
**1. Introduction**

This project had me working with Kesava Asam under Brad Aouizerat to perform a scRNA-seq analysis of pancreatic islet data from patients under differing diabetic status. A recent publication by Wang et al. [1], of which this data was sourced, investigated beta cell dysfunction in type 2 diabetes. Thus, I was directed and guided to obtain the data from this experiment and perform an integrated analysis on the beta cells, utilizing Harmony for integration and SingleR for cell type identity annotation. From this data, we were able to perform a differential gene expression analysis.

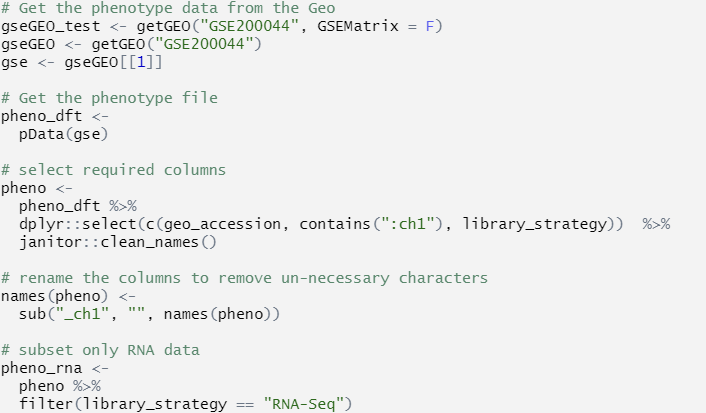
**2. Materials and Methods**



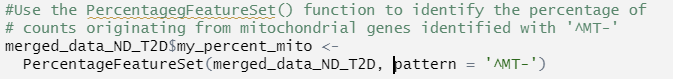
The flowchart above lists the steps taken and key non-Seurat tools used to obtain, process, and analyze the scRNA-seq data. This data was obtained from the GEO page for GSE200044 [2].



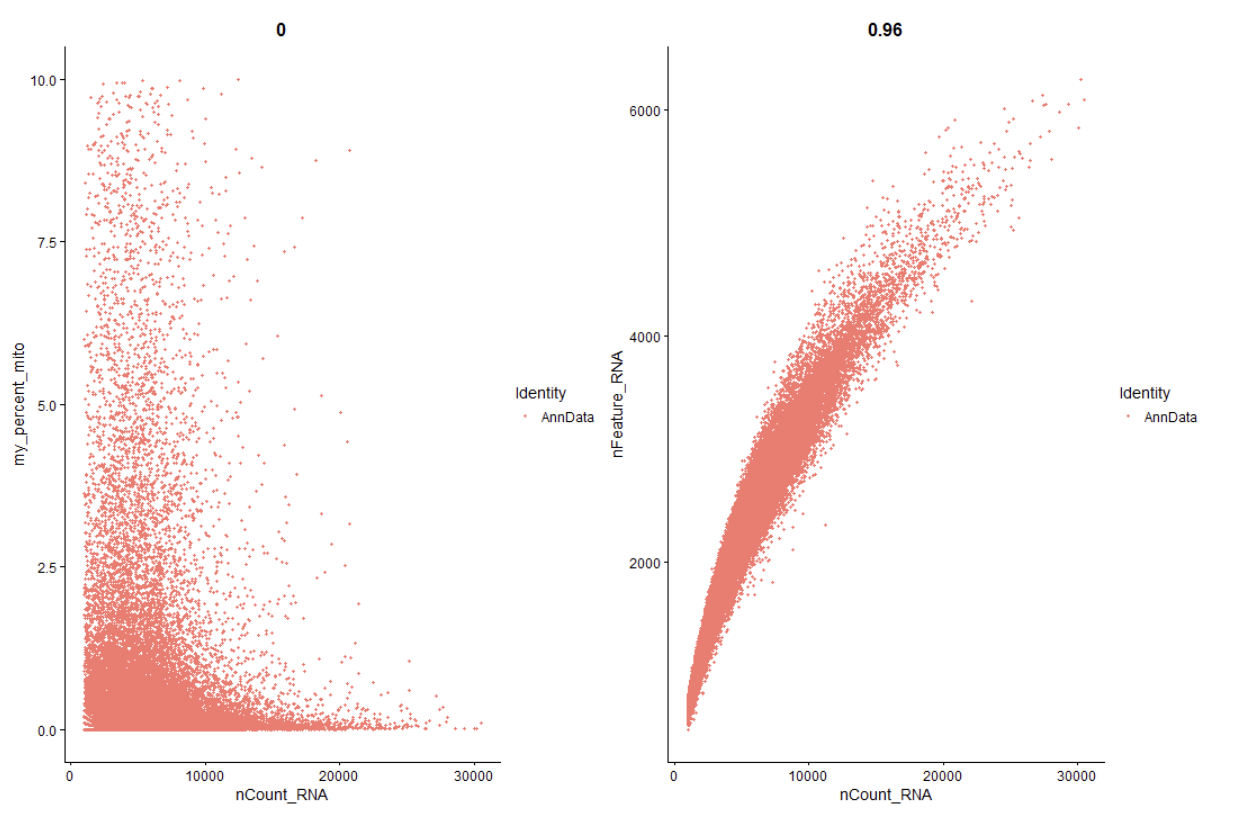


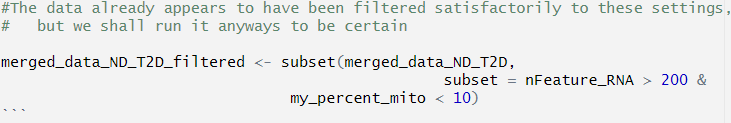


This project was completed in its entirety in R using RStudio, utilizing RMarkdown to create a document consisting of numerous code blocks, enabling me to have additional control over how my code is run and displayed. SeuratDisk[3] and LoadH5Seurat were used to prepare the h5ad file from GEO into a form usable by Seurat, which was then loaded into a Seurat object and had phenotype data merged with the existing data by utilizing **getGEO()**. These steps helped me familiarize myself with Seurat objects as well as the multifaceted h5 (Hierarchical Data) file format, which I had never seen before.

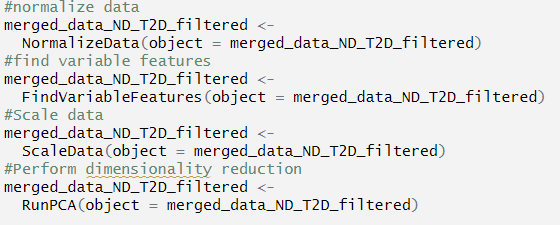


The first processing step in this project was the use of **PercentageFeatureSet()** to identify the percentage of mitochondrial transcripts, since high quantities of it often indicates that these cells are stressed and low quality. [4] This enabled me to plot graphs including mitochondrial and nFeature\_RNA data, for use in determining filtering cutoffs.





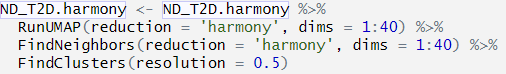
The generated plots indicated that the data was seemingly at least partially processed already, with no obvious outliers and a maximum %MT of 10%. The threshold for mitochondrial filtering is typically 5%, but 10% was chosen here both in agreement with the original Wang et al. paper as well as suggestion from Osorio et al. [4]. Filtering at the determined nFeature\_RNA and percent\_mito can be seen above.





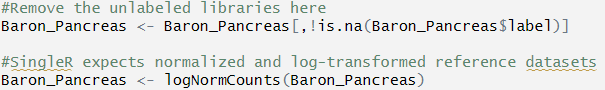
The above code chunks detail the process of determining dimensionality of the dataset, which is accomplished by normalizing and scaling the data while identifying variable features and performing dimensionality reduction. The function **ElbowPlot()** was then used to quickly determine an appropriate cutoff for our number of PCs to use, which was 40.

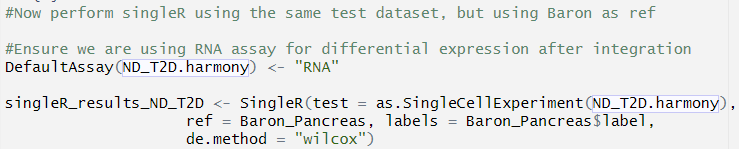




The above code blocks displays the use of **RunHarmony()[5]** to perform integration on the previous object, creating “ND\_T2D.harmony”. This was followed by performing the standard workflow for visualization and clustering, of which several resolutions were tested before choosing to proceed at 1.5, which was also chosen in the original article. Alongside some plots to display the clustering, this completed the integration and provided the top 10 variable features as well, stored in **ND\_T2D.harmony@assays**.







Now that integration and clustering had been performed, **SingleR[6]** was the next step to complete. This tool was used to perform an automated cell type annotation process, propagating the marker gene definition and cluster interpretation from the Baron Pancreas dataset to our own data. After preparing the Baron data and ensuring the RNA assay from the ND\_T2D.harmony was chosen, SingleR was performed with labels stored in **singleR\_results\_ND\_T2D** to be appended to the **ND\_T2D.harmony** metadata. Plotting this updated object grouped by the SingleR labels showed that many of the clusters collapsed into one of fourteen cell types, with alpha and beta cells making up the majority of the cell identities.

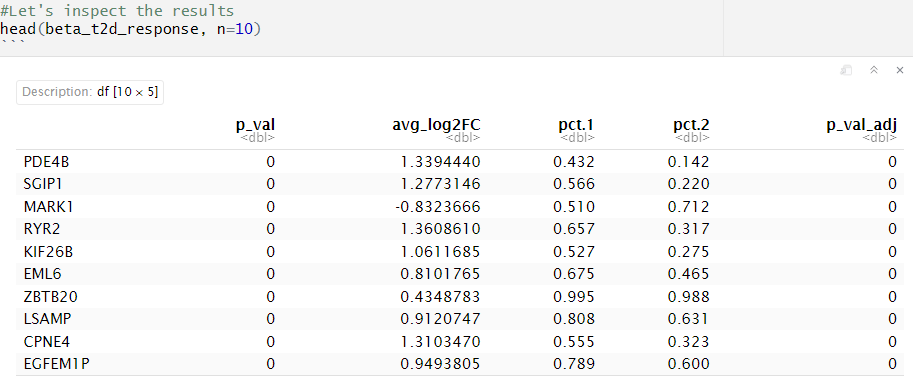






The SingleR labeled data was then modified to create a new column containing both the singleR label and the disease state, enabling me to visualize the ND and T2D conditions for each of the cluster identities, as shown above. Here, the column **cellype.cnd** is added to our existing object with our appropriate labels, which can be visualized using DimPlot after simply setting the **Idents** for **ND\_T2D.harmony** to that same column.





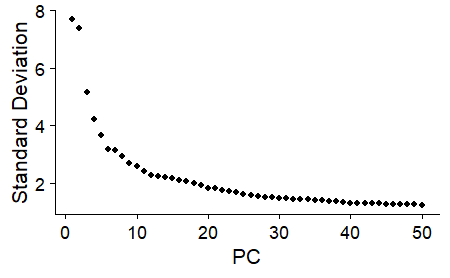
The function **FindMarkers** was then used to find markers/differentially expressed genes for the identity classes using the Wilcoxon Rank Sum test, which in this case were our two disease conditions for the beta cells. Some of the top gene results from this output were then visualized with **FeaturePlot** to see the differential expression between conditions for these genes. This visualization was also expanded beyond these genes to also view the top 3 most positive and negative log2-fold change genes, which may also be of interest. Finally, **VlnPlot** was utilized to provide an alternate way of viewing changes in gene expression, displaying expression level by both conditions for each cell type.

The steps performed in this analysis helped demonstrate to me the capabilities and utility that R Bioconductor tools contained in various R packages can have, increasing the quality and speed of the analytical process with specialized yet powerful tools.

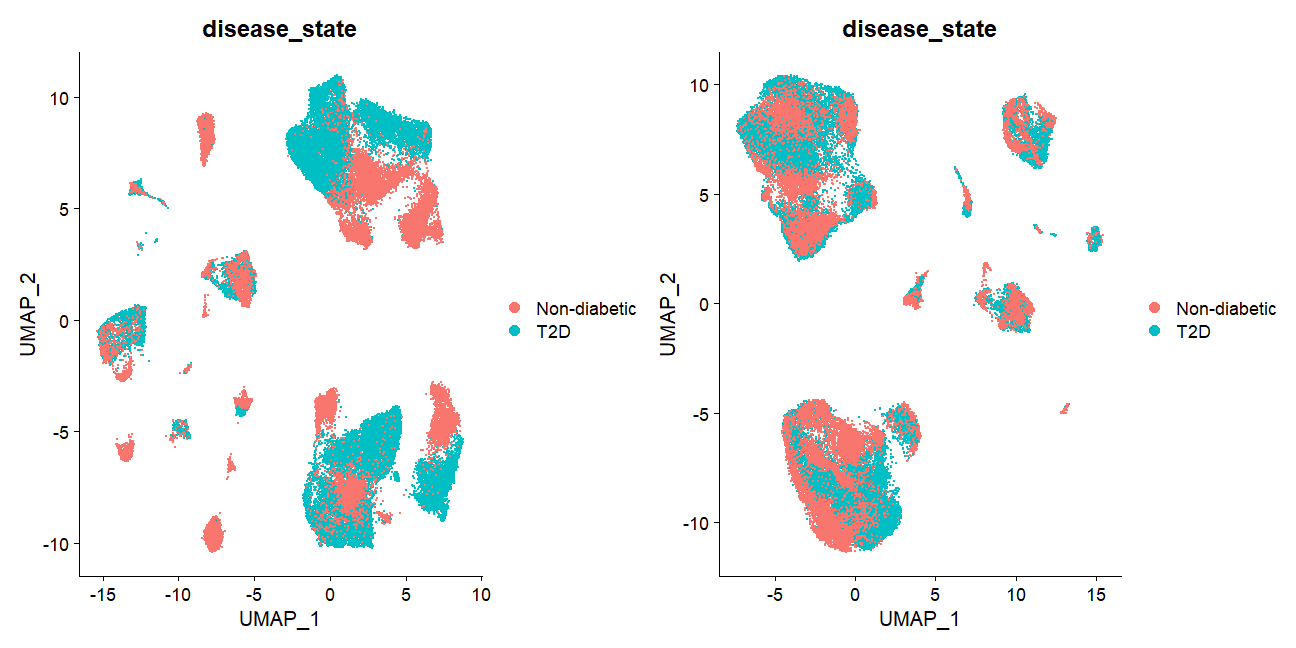
**3. Results and Discussion**

The figures generated from the code blocks in the RMarkdown document provide extremely helpful visualization, illuminating differences between groups or motivating decisions made in the analytical process.

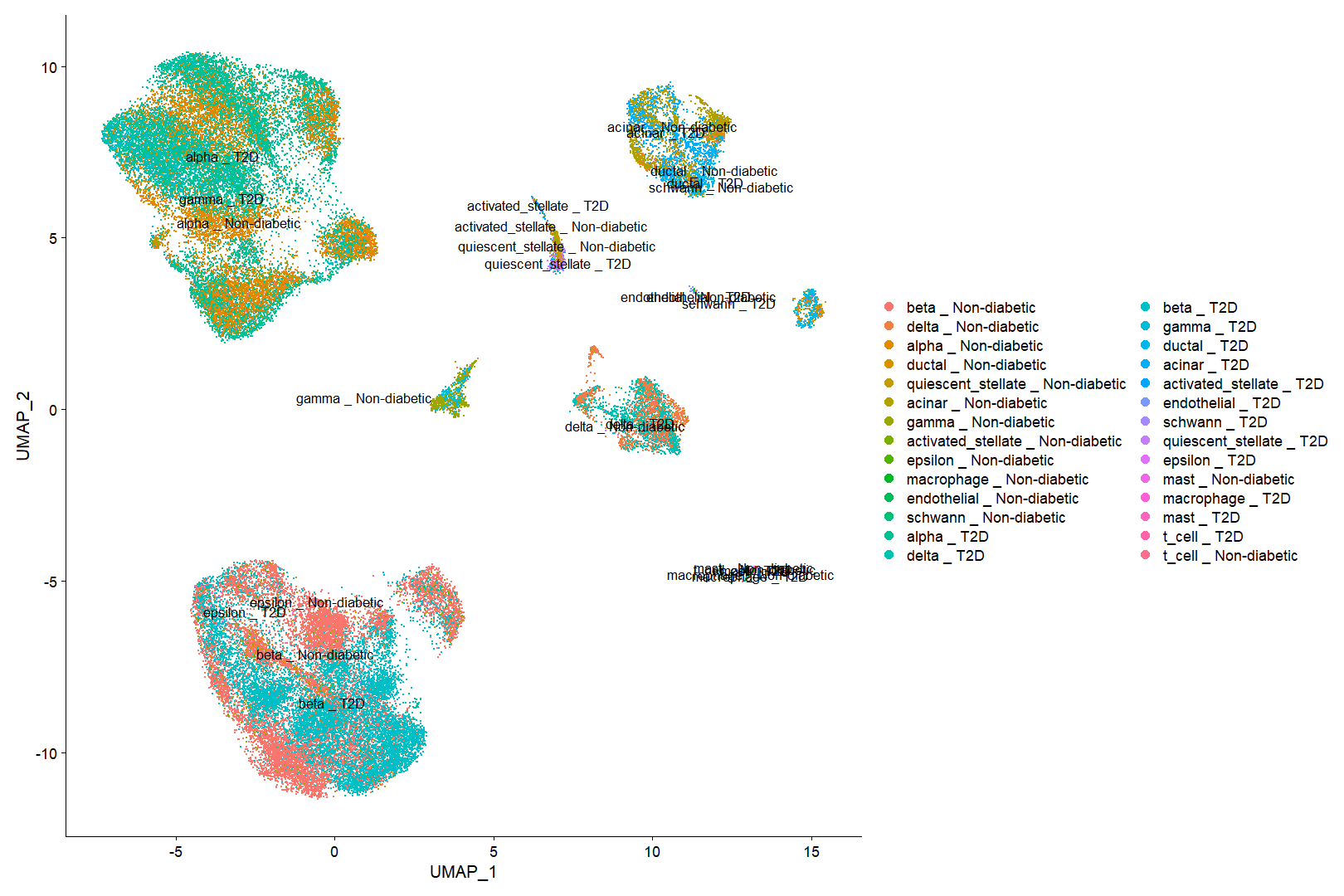
One such plot was the elbow plot, which was used as a tool to determine the dimensionality of the data. While there does appear to be a rough “elbow”around 10-20 PCs, the points continued to gradually decrease in standard deviation until the 30-40s. Of this, I erred on the higher side to capture more rather than less variability if my determination was an underestimate.



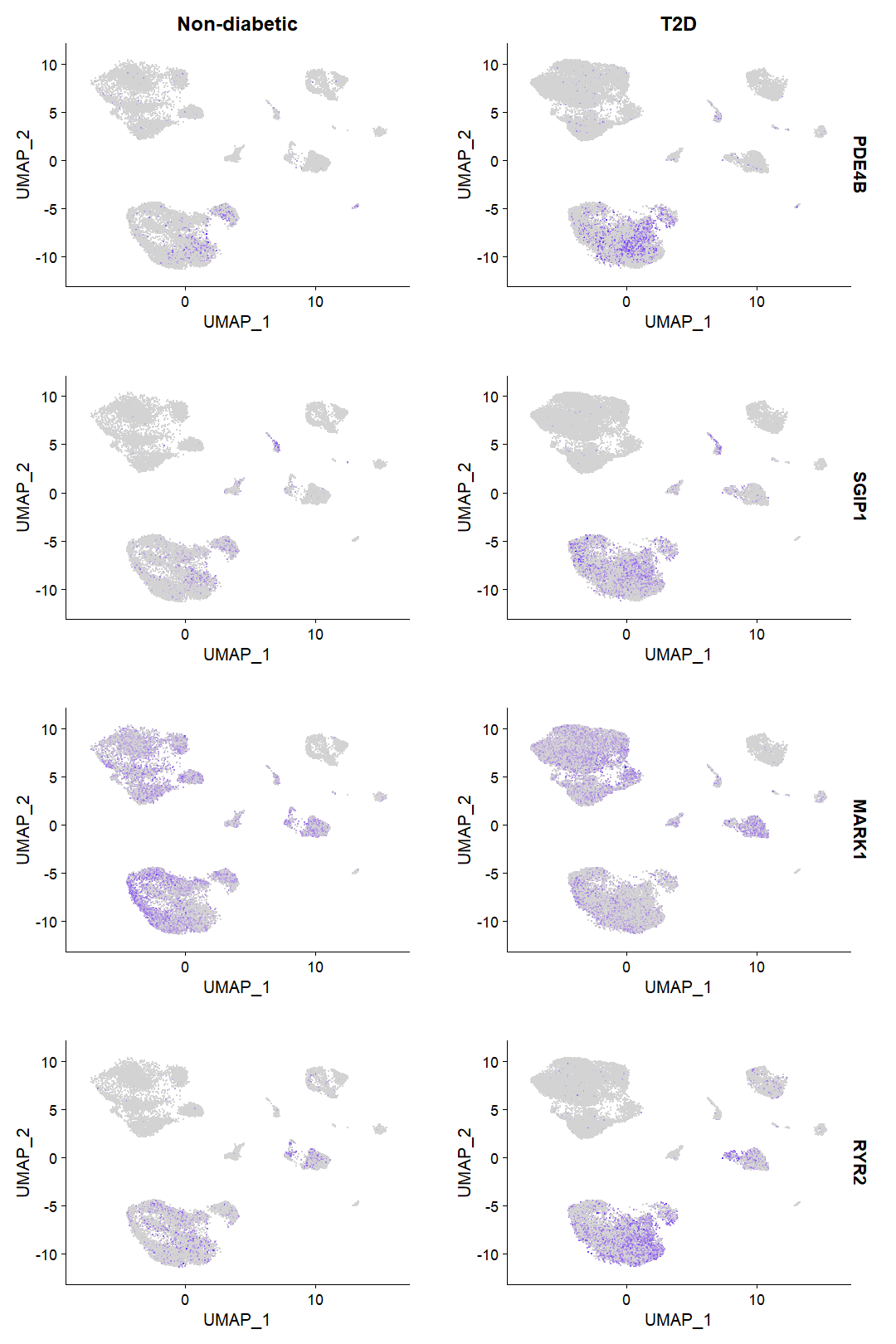
In the following figure, the plots of the integrated data are shown before (left) and after (right) using Harmony to integrate. The image highlights the purpose of integration, ensuring the cell types of one condition (ND) align with the same cell types of other conditions (T2D), and eliminating batch effects from the different samples. We can see that much of the disease\_state separation in clusters is eliminated after utilizing Harmony.



Below, we can see the power of SingleR in its ability to provide quick and robust cell type annotations, provided a good reference dataset is provided, which was the Baron dataset in my case. While alpha and beta cells make up the bulk of the cells, we can see that several cell types are clustering closely together, such as alpha/gamma, beta/epsilon, acinar/ductal, and activated/quiescent stellate cells. This labeling from SingleR allows plotting of the different disease state conditions for the clusters in a follow up plot, which is shown below.

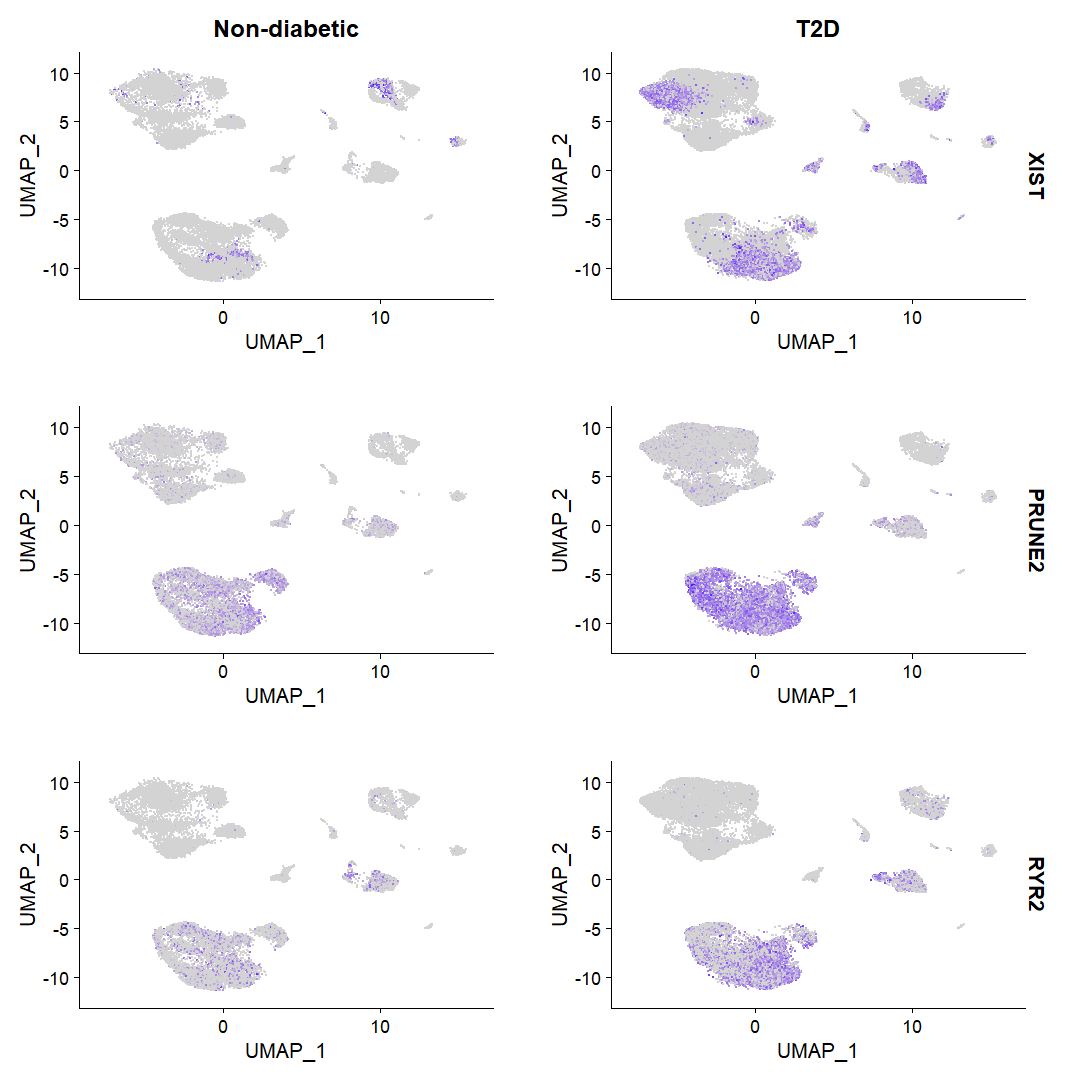


This plot displays the differing disease state conditions for the clusters, with many of the same cell types’ different conditions still grouped closely together. However, it may be of interest to note that for the gamma\_ND cells, we see the label has moved to the previously unlabeled cluster in the center. Since it appears that before splitting by disease state the gamma label is between where we see the ND and T2D labels post-split, there might be some intriguing relationships between the disease state for gamma cells. Although we are proceeding with analysis on the beta cells, this may be of interest in future analysis.

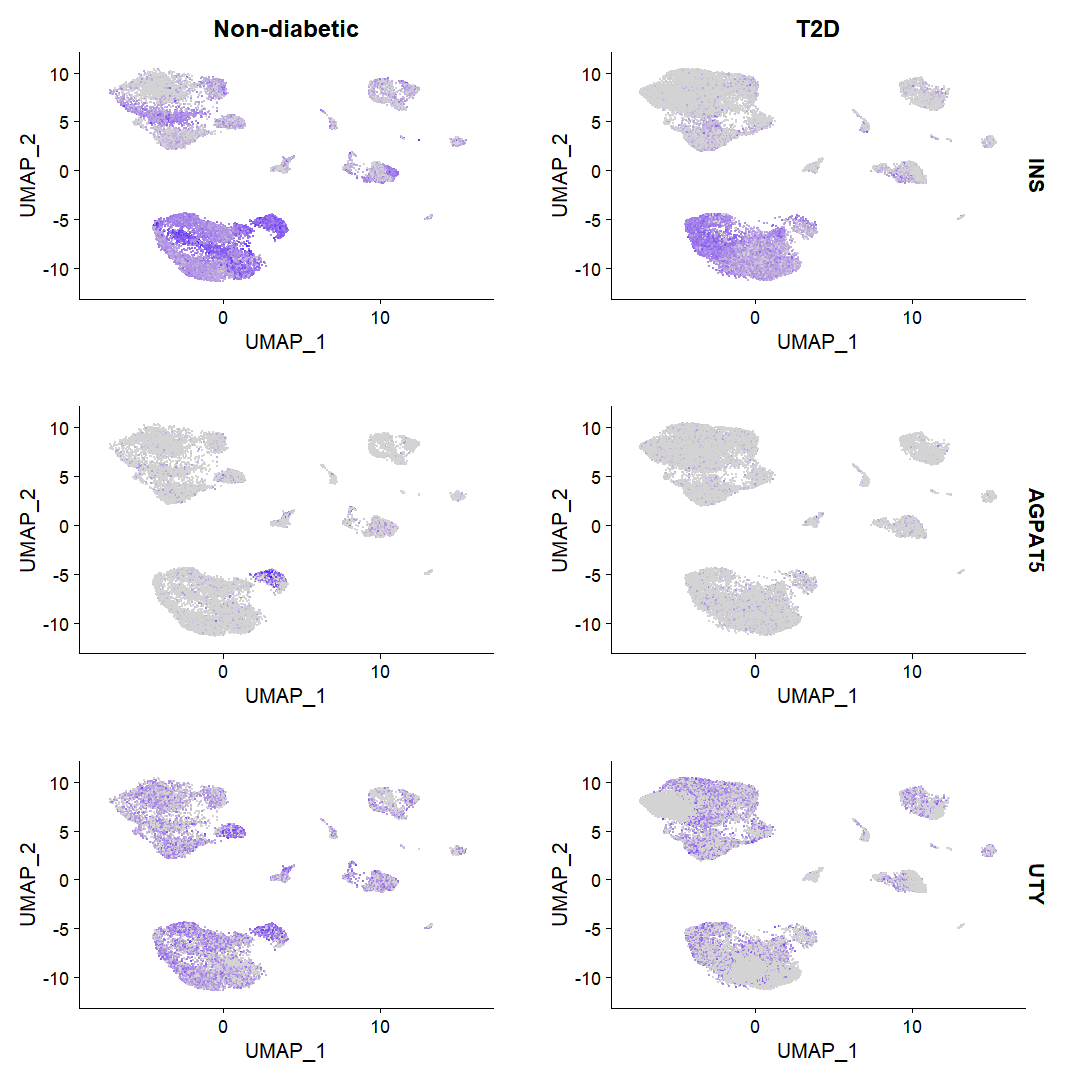


The above image displays FeaturePlots for some of the top genes returned by **FindMarkers** on the ND\_T2D.harmony object, showing an upregulation of PDE4B, SGIP1, and RYR2, and a downregulation in MARK1.

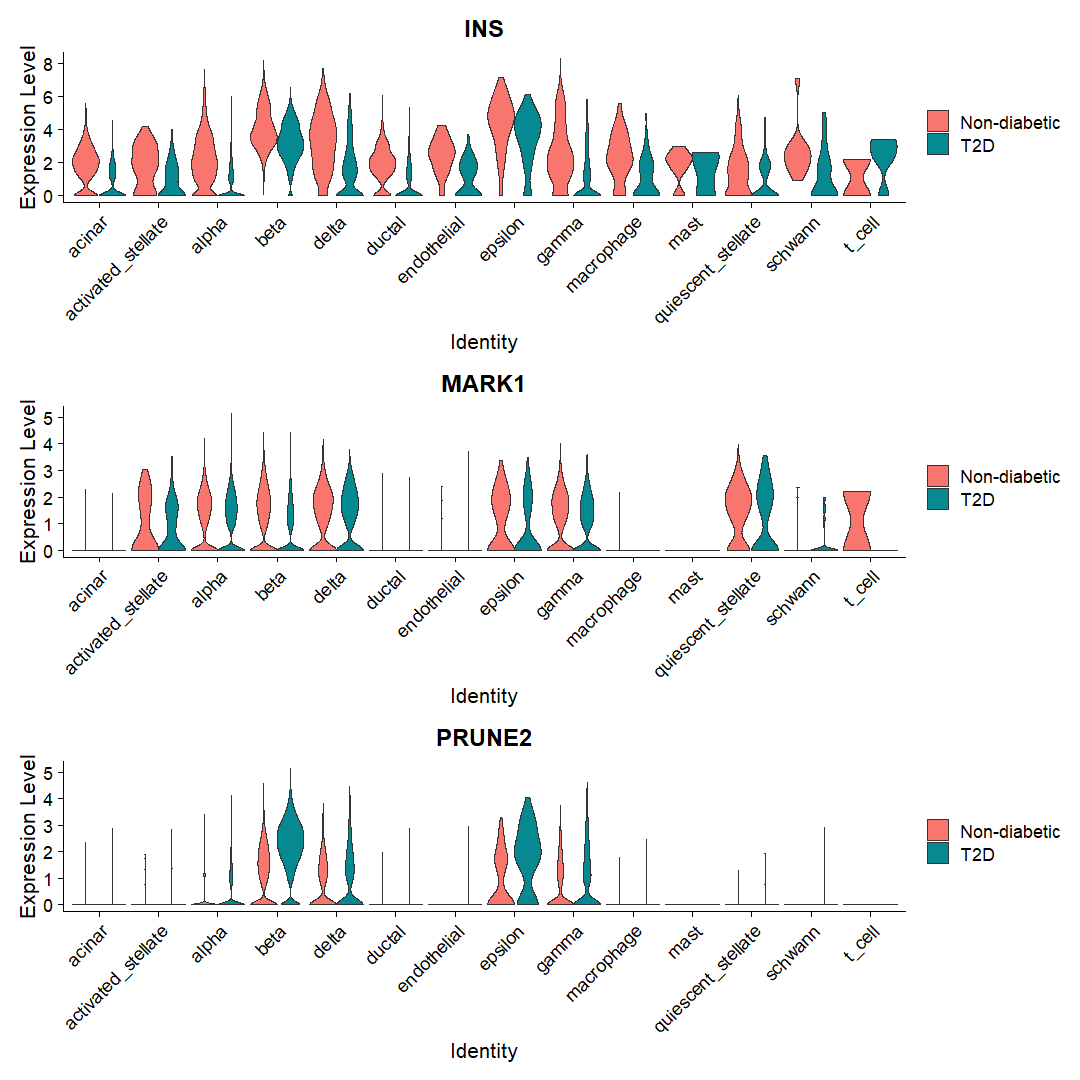
In addition to the top genes returned by FindMarkers, I also inspected the top and bottom avg\_log2FC for the 3 most extreme genes, since this should return the genes with the biggest changes, especially noting that p-values were significant nearly across the board.



The above genes are the most upregulated with log2FC as follows: XIST: 1.718, PRUNE2: 1.393, RYR2: 1.361. We see a large change in XIST, which at first seems odd since XIST is typically only expressed in females as it initiates repression of the inactive X chromosome in female mammals. However, this can be explained by the gender imbalance of the dataset, which is approximately 5:2 = M:F. PRUNE2 encodes a protein that results in reduced stress fiber formation and suppresses oncogenic cellular transformation, especially in prostate cells [7]. RYR2 encodes a ryanodine receptor found in cardiac muscle sarcoplasmic reticulum, and is a component of a calcium channel [8].



The above figure here shows the most down-regulated genes by log2FC, with values as follows: INS: -1.504, AGPAT5: -1.411, UTY: -1.160. We expect to see impaired genes relating to insulin in T2D, so it is reassuring to see that a strong downregulation of the INS gene, which encodes insulin [9], is seen between the disease states. AGPAT5 encodes an integral membrane protein that converts lysophosphatidic acid to phosphatidic acid, the second step in de novo phospholipid biosynthesis [10]. UTY is a Y-chromosome linked gene, and may be differentially expressed due to its potential role in protecting against hypertension [11]. However, because it is Y-linked, this result may not be particularly useful, similarly to how XIST may also suffer from sex imbalance in the original dataset. It may be beneficial to further investigate these genes in future analysis.



Lastly, above you can see several genes from the previous plots can also have their changes in gene expression visualized by violin plot, with those genes being INS, MARK1, and PRUNE2 here. We can see that many of these genes are expressed in multiple cell types, to varying degrees. Especially for PRUNE2, we can more clearly see the expression change specifically for beta and epsilon cells more clearly, which is not as clearly distinguishable in the Feature Plots.

4. **Conclusion**

This project introduced me to performing a scRNA-seq integrated analysis on a large, real, recently published dataset. I also have become more familiar with file formats such as h5/h5ad and modifying Seurat objects as I continuously process the original data. Furthermore, this project acquainted me with several useful and powerful R tools such as Harmony and SingleR to process and annotate the scRNA-seq data more quickly and accurately than what otherwise may be possible alone. By leveraging the use of functions and tools such as those from Harmony, SingleR, and Seurat, I was able to produce helpful plots to better understand the clustering, cell identities, as well as identify differentially expressed genes in beta cells across T2D disease conditions.

In addition to using these tools, this project required a significant amount of modification of data structures such as data frames and Seurat objects, whether that involved merging or adding columns, or modifying objects to prepare them for further downstream analysis. From working with such structures in this project, I better understand how they operate.

In future analysis, it may be of interest to further investigate the gamma cell cluster, as it seems that there is a significant difference between the T2D and ND states for this type. This may further illuminate the effects of T2D on cell types other than beta cells. In addition, because XIST and UTY were some of the genes with the highest upregulations, likely due to gender imbalance in the data, it may be valuable to balance the dataset holistically to see if this significantly affects our results.

**5. References**

1. Wang G;Chiou J;Zeng C;Miller M;Matta I;Han JY;Kadakia N;Okino ML;Beebe E;Mallick M;Camunas-Soler J;Dos Santos T;Dai XQ;Ellis C;Hang Y;Kim SK;MacDonald PE;Kandeel FR;Preissl S;Gaulton KJ;Sander M; “Integrating Genetics with Single-Cell Multiomic Measurements across Disease States Identifies Mechanisms of Beta Cell Dysfunction in Type 2 Diabetes.” *Nature Genetics*, U.S. National Library of Medicine, 23 May 2023, pubmed.ncbi.nlm.nih.gov/37231096/.
2. "The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE200044 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE200044>)."
3. Satija, R; Hoffman, Paul “SeuratDisk v0.0.0.9015” <https://github.com/mojaveazure/seurat-disk>
4. Osorio, D., & Cai, J. J. (2020, August 25). Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8599307/
5. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nature Methods, 16(12), 1289–1296. <https://doi.org/10.1038/s41592-019-0619-0>
6. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M (2019). “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.” Nat. Immunol., 20, 163-172. doi:10.1038/s41590-018-0276-y.
7. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2023 August 18]. Available from: [https://www.ncbi.nlm.nih.gov/gene/15847](https://www.ncbi.nlm.nih.gov/gene/158471)
8. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2023 August 18]. Available from: <https://www.ncbi.nlm.nih.gov/gene/158471>
9. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2023 August 18]. Available from: <https://www.ncbi.nlm.nih.gov/gene/3630>
10. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2023 August 18]. Available from: <https://www.ncbi.nlm.nih.gov/gene/55326>
11. Cunningham, C. M., Li, M., Ruffenach, G., Doshi, M., Aryan, L., Hong, J., Park, J., Hrncir, H., Medzikovic, L., Umar, S., Arnold, A. P., & Eghbali, M. (2022). Y-Chromosome Gene, Uty, Protects Against Pulmonary Hypertension by Reducing Proinflammatory Chemokines. American Journal of Respiratory and Critical Care Medicine, 206(2). https://doi.org/10.1164/rccm.202110-2309oc